Lab 6: Scales

Statistics for Social Scientists II

Bur, GJM

2024-10-14

> *All these considerations justify the vast importance of α in the theory of reliability…It is so pregnant with meaning that it should routinely be applied to all new tests.*
>
> Nunnally and Bernstein (1994): 235

1 Scales generally

Scales, it turns out, are somewhat trivial in themselves. We use them when we want to measure some underlying construct that is hard to get at directly; we get a set of items that measure the underlying construct and then add them up, as if they were questions on a test. For example, we might want to measure someone's overall heart health, but just asking a single question such as "what was your systolic blood pressure today" or "have you ever had a heart attack" could be misleading, so we might want to sum up a series of questions. Or, in another case, someone's attitudes towards a political system might be hard to capture with just one question since respondents may not have full grasp of their true feelings. Someone might, for example, express a high degree of patriotism or pride in one's country but feel that it is "on the wrong track", a common survey response. Asking just "do you love your country" or "do you feel that politicians care about you" would yield opposite results, neither of which would really capture the underlying idea.[1]

So, to construct a scale, you find the items and simply add up all the items. Ideally, your variables are either themselves continuous and on the same scale, or they are dummy variables, since we have established before that dummy variables behave like continuous variables. They have population mean p, where p is defined to be the proportion of individuals who are successes and where *p turns out to be* (for pretty obvious reasons) the mean. We also saw that the variance of a dummy is $p(1-p)$ or in older literature $pq, q := 1-p$. This is because, as we have proved many times (I'll spare you yet another one!), in the population $\sigma^2_Y = \mu_{Y^2} - \mu^2_Y$, and you proved to yourselves in a previous lab meeting and on the homework that this is approximately true in the sample.[2]

[1] In this particular case, we might actually suspect that the underlying construct is *multidimensional*, which makes a single scale problematic, but this is just an illustration of what might be challenging.

[2] Recall that it is not quite true because we divide, in classical statistics, the variance by $n-1$ to make our estimator unbiased, which I think should probably just not be taught anymore until higher levels—it tends to cause people to miss that the variance is *essentially* an average. If we multiply $\overline{y^2} - \overline{y}^2$ by $\frac{n}{n-1}$, which asymptotically approaches unity in large samples, the relation is exact.

1.1 Reliability and validity

It turns out that all of the interesting things to say about scales have to do less with their basic construction than with working out how well they actually measure the underlying construct. We generally break that into two parts, reliability and validity. The reliability of a measure is how well the items in the scale "hang together" or measure something coherent; the validity is how well the scale actually measures the construct in question.

Validity rests on reliability; an unreliable scale, one that essentially measures *nothing*, can't really be a good measurement of any construct, let alone our target.

How do we measure reliability? One fun way to think about doing this, an interesting case of the earliest statisticians coming up with ideas that would only much later get adopted into machine learning when computers got much more efficient, is *split-half reliability*. The idea is that you assign some questions to one half of the test and some questions to the other. One common proposal is to use even-numbered items on a test for half E and odd-numbered items on a test for half O, assuming we have an even number of items. Then, you could take the correlations between all items in half E with all items in half O. If the correlation is high, this would seem to make the test a reliable instrument.

However, there is a problem with this approach, which is that the items could be perfectly correlated *across* halves but have little or no correlation *within* halves—this would mean that the individual halves of the test don't measure anything very coherent. So, one proposal is to try to *average* the correlation between all possible split-halves of the test,[3] and that is one way of defining the well-known *Cronbach's* α (said "alpha"). Don't confuse this with α in the sense of a Type I error!

2 Cronbach's α

It turns out to be easier to derive Cronbach's alpha (α) in a different way. We should instead think of the scale as, in the population, the sum of the random variable that comprise it. Then, the variance of these (highly-correlated) variables is a function of the variances of the individual items and the covariances of all the possible pairs of items. *We will think about α as the ratio of the joint variation, or sum of item covarainces, over the total variance.*

This turns out to have a very nice picture associated with it. Here, below, is the variance-covariance matrix for three random variables, Y_1, Y_2, Y_3. Remarkably, the variance of a variable $Y = Y_1 + Y_2 + Y_3$ —proved below—is equal to the sum of all of the items in the matrix.

$$\begin{bmatrix} \sigma^2_{Y_1} & \sigma_{Y_1,Y_2} & \sigma_{Y_1,Y_3} \\ \sigma_{Y_2,Y_1} & \sigma^2_{Y_2,Y_2} & \sigma_{Y_2,Y_3} \\ \sigma_{Y_3,Y_1} & \sigma_{Y_3,Y_2} & \sigma^2_{Y_3,Y_3} \end{bmatrix}$$

[3] With an even number of items on the test, this is equal to $\binom{k}{\frac{k}{2}}\frac{1}{2}$, or $\frac{k!}{2(k-\frac{k}{2})!(k-\frac{k}{2})!} = \frac{k!}{2(\frac{k}{2})!^2}$. For example, with a small number of items, $k = 4$, we have $\frac{4!}{2 \cdot 2!^2} = \frac{4 \cdot 3}{4} = 3$. You can manually verify this: if our items are labeled $1/4$, the possible groupings for half one are $(1,2), (1,3), (1,4)$, and we don't worry about the other half because each one is already acccounted for; e.g. with $(1,2)$, the other half must be $(3,4)$. That's why we divide by two (in general, if you want to find the number of ways to *partition* a set of things into two groups of size m and $n-m$, it is one half of the binomial coefficient for the number of ways to *pick* m items from a set of n. With more splits, it becomes the more complicated *Stirling number of the second kind*.

2.1 Variances of sums of random variables

An important fact about the variance of composite variables in the population is the following. First, note that the average of a sum of k variables $Y : Y = Y_1 + Y_2 + ...Y_k$ that are equally-weighted (i.e., we just sum each one up to get the total) is …

$$\begin{aligned}
\mu_Y &= \frac{1}{N}\sum_{p=1}^{k}\sum_{j=1}^{N} Y_{pj} \\
&= \sum_{p=1}^{k} \frac{\sum_{j=1}^{N} Y_{pj}}{N} \\
&= \sum_{p=1}^{k} \mu_{Y_p}
\end{aligned}$$

Then, the variance of Y turns out to be the sum of the variances and all of the covariances.

$$\begin{aligned}
\sigma^2_Y &= \frac{1}{N}\sum_{j=1}^{N}(Y_j - \mu_Y)^2 \\
&= \frac{1}{N}\sum_{j=1}^{N}\left\{Y_{1j} + Y_{2j} + ...Y_{kj} - \mu_{Y_1} + \mu_{Y_2} + ...\mu_{Y_k}\right\}^2 \\
&= \frac{1}{N}\sum_{j=1}^{N}\left\{Y_{1j} - \mu_{Y_1} + Y_{2j} - \mu_{Y_2} + ...Y_{kj} - \mu_{Y_k}\right\}^2 \\
&= \frac{1}{N}\sum_{j=1}^{N}\sum_{p=1}^{k}(Y_{pj} - \mu_{Y_p})^2 + \frac{1}{N}\sum_{j=1}^{N}\sum_{p=1}^{k}\sum_{q\neq p}^{k}(Y_{pj} - \mu_{Y_p})(Y_{qj} - \mu_{Y_q}) \\
&= \sum_{p=1}^{k}\sigma^2_{Y_p} + \sum_{p=1}^{k}\sum_{q\neq p}^{k}\sigma_{p,q}
\end{aligned}$$

The last line is often written as follows, since each unique covariance appears twice (we can also drop the Y subscript if it's clear from context).

$$\sum_{p=1}^{k}\sigma^2_{Y_p} + \sum_{p=1}^{k}\sum_{q\neq p}^{k}\sigma_{p,q} = \sum_{p=1}^{k}\sigma^2_p + 2\sum_{p}\sum_{q>p}\sigma_{p,q}$$

Then, one fairly obvious way to measure how much of the variation in a composite variable is due to *shared* variation is to simply divide the total of the covariances by the the total variance, or equivalently to take one minus the idiosyncratic variances.

$$\begin{aligned}\alpha &\approx \frac{\sum_p \sum_{p \neq q} \sigma_{p,q}}{\sigma_Y^2} \\ &\approx \frac{\sigma_Y^2 - \sum_p \sigma_p^2}{\sigma_Y^2} \\ &\approx 1 - \frac{\sum_p \sigma_p^2}{\sigma_Y^2}\end{aligned}$$

So, imagine taking the main diagonal, highlighted here in red, and adding up all of the variances, then dividing by the sum of the whole matrix. Subtracting this number from 1 gives the share of the variance explained by the covariance.

$$\begin{bmatrix} \sigma^2_{Y_1} & \sigma_{Y_1,Y_2} & \sigma_{Y_1,Y_3} \\ \sigma_{Y_2,Y_1} & \sigma^2_{Y_2,Y_2} & \sigma_{Y_2,Y_3} \\ \sigma_{Y_3,Y_1} & \sigma_{Y_3,Y_2} & \sigma^2_{Y_3,Y_3} \end{bmatrix}$$

Equivalently, we could just sum up the covariances directly and divide by the total, though this is maybe more difficult when working by hand.

2.2 The $\frac{k}{k-1}$ adjustment factor

Why the "approximately"? Here is an easy way to think about what's slightly wrong with this formula. Start with the first formulation. Imagine that we swap out covariances for correlations by standardizing each variable before conducting the procedure above; imagine, as well, that all the variables are perfectly correlated. This would cause the numerator to be simply be $k^2 - k$ since there are k^2 possible correlations, but k of them are the correlations of an item with itself. The denominator, however, would just be k^2 since each item has a variance of 1, each pair has a covariance of 1 (since correlations are presumed to be all equal to 1 and the covariance of standardized variables is the correlation), and we have k of the former and $k(k-1)$ of the latter $= k^2$. So, our scale would give $\frac{k-1}{k}$, and we would multiply by $\frac{k}{k-1}$ to get it to be exactly 1. Thus…

$$\begin{aligned}\alpha &= \frac{k}{k-1} \frac{\sum_p \sum_{p \neq q} \sigma_{p,q}}{\sigma_Y^2} \\ &= \frac{k}{k-1} \left\{ \frac{\sigma_Y^2 - \sum_p \sigma_p^2}{\sigma_Y^2} \right\} \\ &= \frac{k}{k-1} \left\{ 1 - \frac{\sum_p \sigma_p^2}{\sigma_Y^2} \right\}\end{aligned}$$

Older texts often use some nice notational simplifications here (Cronbach 1951). So, if you agree to call C_t the total of the *unique* covariances between items (remember that each one is counted twice) or V_t the sum of all the unique variances and covariances we have…

$$\alpha = \frac{k}{k-1}\frac{2C_t}{V_t}$$

So, let's practice a bit. The GSS asks respondents to complete a vocabulary test, which we'll use since it matches the classical test framework, but you could use any variables. I'll show how to do this by hand first. We'll want to drop all missing values here because otherwise our hand calculation will be wrong. Recall that writing something like `d word*` will `describe` all variables starting with the string "word" followed by any number of characters.

```
d word* // let's check out our variables
* Now we write a loop to drop missings on any variable beginning with "word"
foreach v of varlist word* {
    drop if missing(`v')
    }
* Let's make our scale. We just add up items
gen wordscale = worda + wordb + wordc
* Let's get its total variance in working memory
* so that we don't need to find it by hand
sum wordscale, d
local vy = r(Var) // store it in a local
corr worda wordb wordc, cov
di 3/2 * (1 - (r(C)[1,1] + r(C)[2,2] + r(C)[3,3])/`vy')
local handalpha = 3/2 * (1 - (r(C)[1,1] + r(C)[2,2] + r(C)[3,3])/`vy')
alpha worda wordb wordc // Stata's automated command
assert round(r(alpha), .000001) == round(`handalpha', .000001)
scatter wordscale educ, jitter(10) || lfit wordscale educ, legend(off)
```

2.3 Standardizing and Cronbach's α as the Spearman-Brown prophecy formula

Now, in most cases, our items won't be perfectly correlated. But, should we generally standardize them? The answer is that it is often a good idea. The effect of changing a single variable's scale generally requires working out a fairly nasty derivative, but there are also good theoretical "psychometric" reasons to want the items to have the same scale. Empirically, you'll often notice very large increases in α that come from standardizing. In general, the idea is that if you don't standardize your items, items with a much wider range of variation will, if nothing else, have a large impact on your scale.

For example, imagine that we multiply the scale of one of our items by a factor v and let $w = v - 1$. Write $V_{t'}$ for the old sum of item variances and $C_{t'}$ for the old sum of item covariances. Then, we have…

$$= \frac{V_{t'} + w^2\sigma_a^2}{V_{t'} + w^2\sigma_a^2 + C_{t'} + w\sum_{p\neq a}\sigma_{p,a}}$$

Now, working out whether or not toggling w will increase or decrease the ratio requires that tedious (if conceptually straightforward) calculus, but one thing that *is* clear is that this also gives our larger items too much "say"—if w is, say, 1000, terms involving it could of course easily dominate the other items.[4] **So, standardizing is a good idea, often!**

Interestingly, if we keep things standardized, we get a very simple formula for α. The reason is that all of the standardized variances are 1 and all of the covariances are correlations. So, the sum of the correlations is just $k(k-1)\bar{\rho}$ (since there are $k(k-1)$ between-item correlations), the sum of all of the variances is k, and thus the numerator of the "uncorrected" fraction is $k\bar{\rho}$ and the denominator is $k(k-1)\bar{\rho}+k$. Thus…

$$\begin{aligned}\alpha_{std.} &= \frac{k}{k-1}\cdot\frac{k(k-1)\bar{\rho}}{k+k(k-1)\bar{\rho}} \\ &= \frac{k\bar{\rho}}{1+(k-1)\bar{\rho}}\end{aligned}$$

This last form is called the Spearman-Brown prophecy formula, which is a great name. It tells us that α is a function of the correlations between pairs of items and the number of items. All else equal, i.e. assuming that the average item correlations don't change, as the number of items increases, the test becomes more reliable. And, of course, holding the items constant, as the average correlation increases, we have a better scale.

Let's try making some standardized variable and then applying it to our scale. Here, we should not expect the value to change too much since each variable is already a dummy.

```
* standardize the variables

local scalevars "worda wordb wordc"
    * Note that this just puts the names of the variables into a
    * string. You can even go more minimal without the quotation
    * marks, but this is safer generally.

* Making standardized versions of our variable with a loop
foreach var of varlist `scalevars' {
    sum `var'
    gen z_`var' = (`var' - r(mean))/r(sd)
    }
* Note in what follows that z_* is a wildcard that will just put all
    * of the variables starting with that phrase into the variable list
    * Obviously, use with caution, but it is rare to have data-sets with
    * that type of name for non-standard variables. If you do, just rename
    * as appropriate when standardizing.
local stdvars z_*
alpha `stdvars'
alpha `scalevars', std
```

[4] You might see some sources write that *increasing* the total variance of the scale increases the reliability, as in (Nunnally and Bernstein 1994: 237), but that is only true *holding* the individual item variances constant (since by algebra that must mean the covariances increase): "If, for example, two 20 item tests *have the same average [variance]*, the one with the larger variance is more reliable" (emphasis mine).

We can quickly verify that we can think about this conceptually in very simple terms, using the Spearman-Brown formula, as a function of the correlations and the number of items ("length of test").

```
* Spearman-Brown
local corrs = 0
    * we will successively sum the interitem correlations
    * but we need to initialize a loop.

* Note that we have a double-loop here, which maybe helps
* those double sums before seem less scary. We just start
* with one value of p, run the whole sum over q, then go to
* the next value of p and keep adding! We will update the
* local as we go by redefining it within the loop to be the
* existing value plus the value

* Note that we need an -if- statement to exclude cases where
* we _would_ take the correlation of a variable with itself!

foreach p of varlist z_* {
    foreach q of varlist z_* {
        if `p' != `q'{
            corr `p' `q'
            local corr = r(C)[2,1]
            local corrs = `corrs' + `corr'
            di "The corr between `p' and `q' is `corr'"
            }
        }
    }
di `corrs'
local rbar = `corrs'/(2*`k') // we counted each one twice.
di `rbar'
local sbpf = (`k'* `rbar')/(1 + (`k'-1)*`rbar')
di `sbpf'
alpha `scalevars', std
assert round(`sbpf', .000001) == round(r(alpha), .000001)
```

3 Various other issues

3.1 Reverse scoring

There are various other practical issues that we need to mention. First, our reliability coefficient assumes that items have the same *direction* of association—if we have a bunch of items that are all very correlated but in opposite directions, α could even end up being zero! Stata's `alpha` command generally will automatically infer the proper direction and reverse some items, but sometimes it has trouble doing so. You can manually specify this with `alpha x1 x2 ... xk, reverse(xi xj)`. However, you can also just manually reverse score your variables very easily with `gen xirev = -1*xi`, etc.

3.2 Cutoffs for α

This is just as arbitrary as the other α we know, in fact! So, I'll just tell you that the conventional wisdom is to seek $\alpha \geq 0.7$. (Tavakol and Dennick 2011) report a wide range of values considered acceptable, from 0.7 to 0.95.

3.3 Sampling theory

I'll largely omit that from these notes. As (Nunnally and Bernstein 1994: 228) note, it can become very complicated:

> There is a double problem of sampling related to the precision of reliability estimates—the sampling of > objects (usually people) and the sampling of items … it is very difficult to consider both sampling problems simultaneously.

The "sampling of items" refers to the fact that you can choose to include different variables in your scale, so there is a kind of "second sampling" on top of the fact that your observations are themselves the product of a random sample. You can bootstrap α if need be.

4 Exercise

Find α by hand and using the automated routine for all of the items starting with `word`. The GSS already makes a scale for you, just called `wordsum`, but .

5 Other perspectives on what α means

Many of you who are largely qualitative scholars might be interested in *inter-rater reliability*. I'll just mention that it is conceptually very similar to α; see (DeVellis 2003).

We can also interpret a single split-half reliability, the (average or total) correlation of all pairs of items where the first item is in half a and the second is in half b, as the an estimate of the *squared* correlation between a single test and the true, unobserved score; if we had an infinite number of tests (God help us all), the true score, per the *domain sampling model*, would be nothing other than that score.[5] So, Cronbach's α is the average of such estimates and can be interpreted as a guess at the share of variance in the true score explained by our noisy measures, and its root is an estimate correlation of our scale with any other scale.

6 Appendix

6.1 Cronbach's α as the average of split-half reliabilities

First, define the general split-half reliability as follows. I'll call this r_{tt} following (Cronbach 1951); don't confuse this with the regular correlation coefficient. Call C_a the total covariance between items within half a, C_b the same but for half b, and $C_{a,b}$ the covariance between all items in half a with all in half b; the total covariance matrix then has four parts, as illustrated below. Let C_t denote the total of all of the covarainces; $C_t = 2(C_a + C_b + C_{a,b})$.

[5] Some other theories posit that the true score is literally unobservable even with infinite tests).

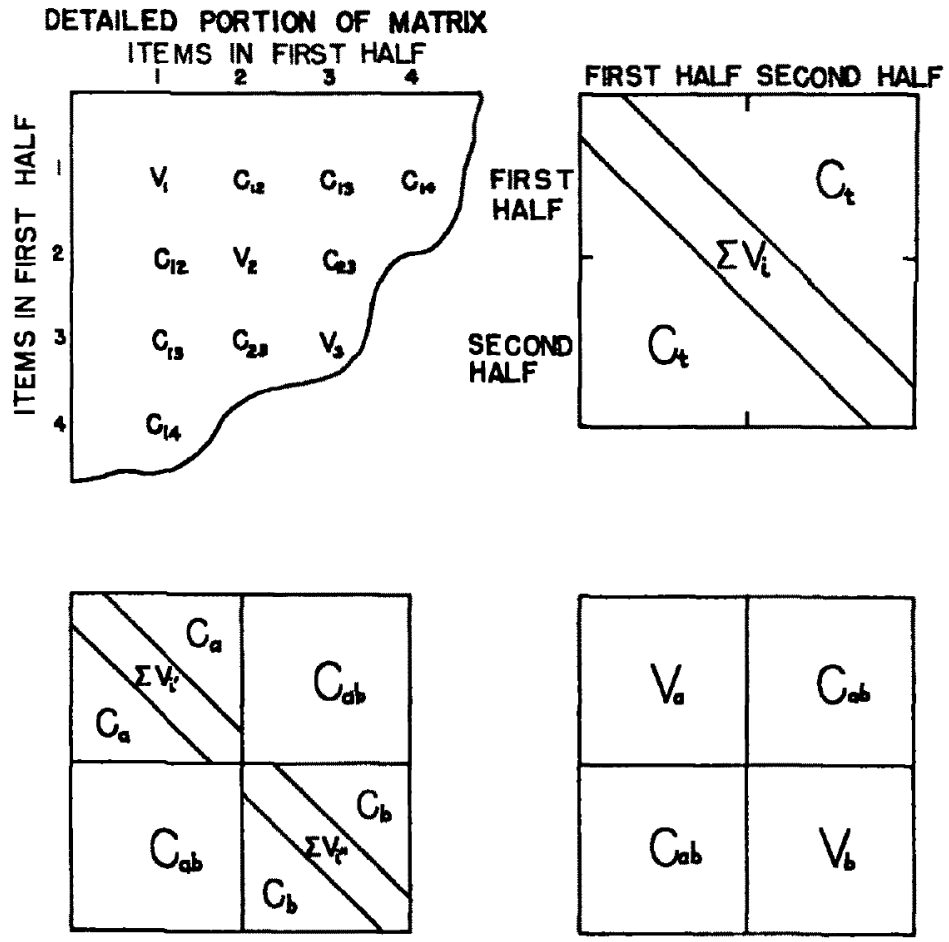

FIGURE 1
Schematic Division of the Matrix of Item Variances and Covariances.

Figure 1: Cronbach's original diagrams

Then, if we make the assumption that $2C_a = 2C_b = C_{a,b}$, we can write a formula for the total share of the variance attributable to the covariances as…

Our reliability measure is the sum of the covariances shared across tests

$$r_{tt} = \frac{2(C_a + C_b + C_{a,b})}{V_t}$$
$$r_{tt} = \frac{4C_{a,b}}{V_t}$$

Now, we can write the average of the covariances as $\overline{C}_{p,q} = C_t \frac{1}{k(k-1)/2}$, and we can rewrite α as $\frac{k^2 C_t}{V_t}$. We can also write the average of all covariances across split-halves as $\bar{r}_{tt} = \frac{4\overline{C}_{a,b}}{V_t}$

Next, note that the probability that any particular covariance falls within a given split half is the number of possible covariances between halves, $(\frac{k}{2})^2$, over the number of possible covariances $\binom{k}{2} = \frac{k(k-1)}{2}$. Dividing yields $\frac{k}{2(k-1)}$.

So, to sum over all possible split-half covariances $\sum C_{a,b}$, we multiply the number of such splits by the probability of seeing some particular between-half covariance in that split. Let $m = k/2$ for simplicity so that the number of splits is $\binom{k}{k/2} = \frac{k!}{2(k/2)!^2}$. Then we have…

$$\sum C_{a,b} = \frac{k!}{2(k/2)!^2} \frac{k}{2(k-1)} \sum_{p \neq q} C_{p,q}$$

Since the sum of the covariances is also $\frac{k(k-1)}{2} C_t$, we can combine and write…

$$\sum C_{a,b} = \frac{k!}{2(k/2)!^2} \frac{k}{2(k-1)} \frac{k(k-1)}{2} \bar{C}_{p,q}$$
$$\overline{C}_{a,b} = \frac{k^2}{4} \overline{C}_{p,q}$$

Finally, since $\bar{r}_{tt} = \frac{4\bar{C}_{a,b}}{V_t}$

$$\bar{r}_{tt} = \frac{k^2 \overline{C}_{p,q}}{V_t} = \alpha$$

6.2 α as an estimate of the degree to which our scale explains variation in true scores

First, write out the correlation between any item, say item 1, and all other items in the test as follows. Assume that the variables are standardized, let the items be called $z_1, z_2, ...z_k$, and assume all sums without indices are sums over people.

$$
\begin{aligned}
r_{1,(2\to k)} &= \frac{\sum z_1(z_1, z_2, ...z_k)}{\sqrt{\sum z_1^2}\sqrt{\sum (z_1 + z_2 + ...z_k)^2}} \\
&= \frac{1 + r_{1,2}... + r_{1,k}}{1\sqrt{k + k(k-1)\bar{r}_{ij}}} \\
&= \frac{1 + (k-1)\bar{r}_{ij}}{\sqrt{k + k(k-1)\bar{r}_{ij}}}
\end{aligned}
$$

Then, divide both numerator and denominator by k. As k gets tolerably large, we simply end up with $r_{1,(2\to k)} = \frac{\bar{r}_{ij}}{\sqrt{\bar{r}_{ij}}} = \sqrt{\bar{r}_{ij}}$. This is the correlation of a given item with the whole test since making k arbitrarily large makes those items the true score by definition under the domain sampling model—or, crucially, *any* test if we assume interitem correlations to be very similar. In words, the correlation of a given item with the true test is equal to the root of the average inter-item correlation. (Nunnally and Bernstein 1994: 221) call the interitem correlation the reliability *coefficient* (for individual items or for all; they assume them equal) and write it $\bar{r}_{ij}$; they call the correlation of any item with the true score to be the reliability *index* and denote it r_{it}. The relation is that $r_{it} = \sqrt{\bar{r}_{ij}}$.[6]

Now, we simply assume that each z-score is a score on a whole test rather than an individual item. Nothing really changes, except that it is now *more* plausible that tests have approximately the same correlation.

Next, since $r_{it}^2 = \bar{r}_{ij}$, this means that $\bar{r}_{ij}$ can be interpreted as the square of some particular kind of correlation coefficient with all the attendant meaning; so, it can be interpreted, among other ways, as the share of variance in the unobserved true measure explained by the noisy measures if we could observe the true measure and regress it upon the noisy measures.

6.3 Critiques of α

In psychometric theory, the concept has received quite a bit of criticism. See (McNeish 2017) for a review. Some of this is due to the fact that Cronbach's originally-cited article appears to be in the top 100 cited articles *period* in English (McNeish 2017: 2), so there is quite a strong incentive to "fact-check" it. I'll briefly discuss the four assumptions McNeish holds are usually not met.

1. τ-equivalence ("tau equivalence")

   This refers to the assumption that the effects of each item on the true, unmeasured, underlying construct (all terms used synonymously at times) have the same effect on the outcome. Of course, this is usually false. However, this only makes α a conservative estimate of reliability, which is generally not a problem at all. An alternative approach is to find a coefficient called ω (*omega*), which uses the assumption of different *factor loadings*, not discussed here (see Devellis 2003 for an easy overview).

[6] Note by the way that (Nunnally and Bernstein 1994: 221) appears to have a critical typo: the authors refer to both the "[t]he crucial assumption of equal correlations among scoring units $\bar{r}_{1j} = \bar{r}_{ij}$" but also their equation 6-7a states that $\bar{r}_{1j} = \sqrt{\bar{r}_{ij}}$, both of which cannot be true unless both numbers = 1.

2. Items are continuous with normal distributions.

   The continuous part of the assumption assumes that our data are *drawn from* continuous distributions (remember: all data are discrete; continuity is a theoretical property of a population distribution). This is important, but it is often ignored in practice when we have, say, Likert items we think are reasonably well-measured. *One other solution is to take such items and make them binary*; then, as we have learned, the resulting dummy variables have more properties of continuous variables, such as a meaningful average and variance. Another approach is to use what are called *tetrachoric* and *polychoric* correlations, which assume that you observe binary or ordinal but are *censored* forms of continuous variables. However, this is really just a theoretical claim, not a mathematical one; it is possible that your items really are binary or really are ordinal.

   The normality assumption is more important for sampling theory, and there are nonparametric methods (such as the bootstrap) that . **In general, works in applied statistics constantly overstate the need to assume normality** (e.g., you'll often see people say that "correlation" or "regression" require assumptions that the variables are marginally Normal, but this is frankly complete nonsense; we can and have derived the OLS normal equations using nothing but linear algebra with no reference to density curves whatsoever). This is such a cliché that it is something of a cliché to rebut it (Grayson 2004). We typically assume only that sampling distributions for some simple sample statistics are Normal or Gaussian, and *some* ways of deriving the OLS estimators assume that *residuals* are Normal. **Claims that statistics generally presumes Normally-distributed variables is wrong**.

3. Uncorrelated errors.

   This means that we assume that responses to the questions are generally only correlated to the extent that the true scores are correlated, not that things like the placement of items in a survey affects each other (e.g., respondents might be more agreeable at the end of the test just to get the thing over with). This is important because this can cause α to be overstated, but it is a problem for surveys generally.

4. Unidimensionality.

   In short, "large Cronbach's alpha value does not necessarily guarantee that the scale measures a single construct" but instead might be measuring . Again, this is something that is hard to actually test without more complicated techniques such as factor analysis; it is largely a theoretical-analytical question.[7]

References

Cronbach, Lee J. 1951. "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika* 16(3): 297–334.

[7] Nunnally and Bernstein's (1994: 218) point is apposite: "The average correlation in the matrix $\bar{r}_{ij}$ indicates the extent to which a common core exists among the variables. It is not necessary that this core be a single factor in the sense of [factor analysis]. The dispersion of correlations about this average indicates the extent to which variables vary in sharing this common core. If one assumes that all variables share equally in this core, the average correlations in each column of the hypothetical matrix would be the same and would equal the average correlation in the whole matrix $\bar{r}_{ij}$".

DeVellis, R. F. 2003. *Scale Development: Theory and Applications*. 2nd edition. 2nd edition. SAGE Publications.

McNeish, D. 2017. "Thanks Coefficient Alpha, We'll Take It from Here." *Psychological Methods* 23(3): 412–33.

Nunnally, J. C., and I. H. Bernstein. 1994. *Psychometric Theory*. 3rd edition. 3rd edition. McGraw-Hill.

Tavakol, M., and R. Dennick. 2011. "Making Sense of Cronbach's Alpha." *International Journal of Medical Education* 2: 53–55.